Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 11, November-2019

Automatic Recognition, Identifying Speaker Emotion and Speaker Age Classification using Voice Signal

Automatic Age Classification using Voice Signal

Dr. A. S Umesh
Computer Engineering Science
TIT, Bhopal

Prof. Ramesh Patole
Information technology department
G. H. R.C.M, Pune

Prof. Krishna Kulkarni
ICOER, Pune

***Abstract*— Audio engineers commonly refer to recording and reproduction systems as "chains," an apt designation because it invites attention to links. Although digital management of audio information offers improvements in signal quality over analog methods, digital systems are not without problems. Aliasing errors, sampling rate jitter, amplitude distortion, intermodulation distortion, spurious output signals, inter-channel cross-talk, inter-channel phase distortion, idle channel noise, and delay distortion can occur and are the subject of technical standards (Audio Engineering Society, 1998). Because digital audio systems by definition accept and produce analog signals, several purely analog issues remain relevant. Humans are very good at recognizing people. They can guess a person's gender, age, accent, and emotion by just hearing the person's voice over the phone. At the highest level, people use semantics, diction, idiolect, pronunciation and idiosyncrasies, which emerge from socio-economic status, education and place of birth of a speaker. At the intermediate level, they use prosodic, rhythm, speed, intonation and volume of modulation, which discriminate personality and parental influence of a speaker.**

Keywords—Raspberry pi, speaker and audio files; spell check algorithm; pattern matching algorithm; deep learning; machine learning; neural network;

1. INTRODUCTION:

Humans are very good at recognizing people. They can guess a person's gender, age, accent, and emotion by just hearing the person's voice over the phone. At the highest level, people use semantics, diction, idiolect, pronunciation and idiosyncrasies, which emerge from socio-economic status, education and place of birth of a speaker. At the intermediate level, they use prosodic, rhythm, speed, intonation and volume of modulation, which discriminate personality and parental influence of a speaker. At the lowest level they use acoustic aspects of sounds, such as nasality, breathiness or roughness. Recordings of the same utterance of two people will sound different because the process of speaking engages the individual mental and physical systems. Since these systems are different among people, their speech will be also different even for the same message. The speaker-specific characteristics in the signal can be Exploited by listeners and technological applications to describe and classify speakers, Based on age, gender, accent, language, emotion or health.

There are many speaker characteristics that have useful applications. The most popular of these include gender, age, health, language, dialect, accent, socialist, idiolect, emotional state and attention state. These characteristics have many applications in Dialog Systems, Speech Synthesis, Forensics, Call Routing, Speech Translation, Language Learning, Assessment Systems, Speaker Recognition, Meeting Browser, Law Enforcement, Human-Robot Interaction, and Smart Workspaces. For example, the Spoken Dialogs Systems provide services in the domains of finance, travel, scheduling, tutoring or weather. The systems need to gather information from the user automatically in order to provide timely and relevant services. Most telephone-based services today use spoken dialog systems to either route calls to the appropriate agent or even handle the complete service by an automatic system. Some of the reasons for automatic speaker classification include: automatic indexing of audio material, identification or verification of people to ensure secure access, loading pre-trained models for speech recognition tasks, tailoring machine dialogue to the needs and situation of the user, or synthesizing voice with similar characteristics (gender, age, and accent) to the speaker. Demand for human-like response systems is increasing. For example, shopping systems can recommend suitable goods appropriate to the age and sex of the shopper.

Technical key words: -

- SVM Super vector approach
- The sound wave, a band pass filter of bandwidth,
- Vector quantization techniques,
- Recognition Algorithm,
- Mel Frequency Cepstral Coefficient (MFCC) & Vector Quantization (VQ).
- Speaker age, Phonetics, Acoustic analysis, Acoustic correlates

The human voice consists of sounds produced by a human being using the vocal folds for carrying out acoustic activities

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 11, November-2019

such as talking, singing, laughing, shouting, etc. The human voice frequency is specifically a part of the human sound production mechanism in which the vocal cords or folds are the primary source of generated sounds. Other sound production mechanisms produced from the same general area of the body involve the production of unvoiced consonants, clicks, whistling and whispering. Generally, the mechanism for generating the human voice can be subdivided into three parts; the lungs, the vocal folds within the larynx, and the articulators. The human voice and associated speech patterns can be characterized by a number of attributes, the primary ones being *pitch, loudness or sound pressure, timbre, and tone*. *Pitch* is an auditory sensation in which a listener assigns musical tones to relative positions on a musical scale based primarily on their perception of the frequency of vibration . Pitch can be quantified as a frequency, but it is based on the subjective perception of a sound wave. Sound oscillations can be measured to obtain a frequency in hertz or cycles per second. The pitch is independent of the intensity or amplitude of the sound wave. A high-pitched sound indicates rapid oscillations, whereas, a low-pitched sound corresponds to slower oscillations. Pitch of complex sounds such as speech and musical notes corresponds to the repetition rate of periodic or nearly-periodic sounds, or the reciprocal of the time interval between similar repeating events in the sound waveform.

Loudness is a subjective perception of sound pressure and can be defined as the attribute of auditory sensation, in terms of which, sounds can be ordered on a scale ranging from quiet to loud. Sound pressure is the local pressure deviation from the ambient, average, or equilibrium atmospheric pressure, caused by a sound wave. *Sound pressure level* (SPL) is a logarithmic measure of the effective pressure of a sound relative to a reference value and is often measured in units of decibel (dB). The lower limit of audibility is defined as SPL of 0 dB, but the upper limit is not as clearly defined.

Timbre is the perceived sound quality of a musical note, sound or tone. Timbre distinguishes different types of sound production and enables listeners to distinguish different instruments in the same category. The physical characteristics of sound that determine the perception of timbre include spectrum and envelope. Figure 1 shows a sound wave with its temporal envelope marked in red. In simple terms, timbre is what makes a particular sound be perceived differently from another sound, even when they have the same pitch and loudness.

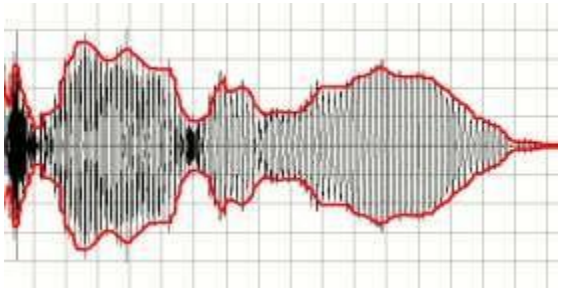

Fig 1: A sound wave's envelope marked in red

Tone is the use of pitch in language to distinguish lexical or grammatical meaning – that is, to distinguish or to inflect words. All verbal languages use pitch to express emotional and other paralinguistic information and to convey emphasis, contrast, and other such features.

It has been frequently observed that the tonal quality of the human voice changes while expressing various emotions. With different emotions and moods, not only does the tonal quality vary, but the associated speech patterns change too. For instance, people may tend to talk in loud voices when angry and use shrill or high-pitched voices when in a scared or panicked emotional state. Some people tend to ramble when they get excited or nervous. On the contrary, when in a pensive emotional state, people tend to speak slowly and make longer pauses, thereby indicating an increase in time spacing between consecutive words of their speech.

Detection of human emotions through voice- and speech-pattern analysis can prove to be beneficial in improving conversational and persuasion skills, especially in those circumstances or applications where direct face-to-face human interaction is not possible or preferred. The practical applications of human emotion detection through voice and speech processing could be numerous. Some such plausible real-world situations could be while conducting marketing surveys with customers through telephonic conversations, participating in anonymous online chat-rooms, conducting business voice-conferences and so on.

This paper presents an algorithmic approach which aids in detection of human emotions, by analyzing various voice attributes and speech patterns. Section II describes the proposed algorithmic approach in detail, along with the obtained results. Section III is devoted to result analysis. Finally, section IV elucidates the conclusions derived from the proposed study and future scope of work.

2. ALGORITHMIC APPROACH TO DETECTION OF HUMAN EMOTIONS

This section describes an algorithmic approach for deducing human emotions through voice- and speech-pattern analysis. In order to achieve this objective, three test cases have been examined, corresponding to the three emotional states: *normal* emotional state, *angry* emotional state, and *panicked* emotional state. For carrying out the analysis, four vocal parameters have been taken into consideration: pitch, SPL, timbre, and time gaps between consecutive words of speech. In order to quantitatively represent timbre, its temporal envelope for advance and decay times has been considered. The primary function of the proposed algorithmic approach is to detect different emotional states by analyzing the deviations in the aforementioned four parameters from that of the normal emotional state. The proposed analysis was carried out with the help of software packages such as MATLAB and Wave pad.

- **Case 1**: *Normal* emotional state

  This test case involves statistics for pitch, SPL, timbre, and word-timing gaps derived from speech samples that were orated while the speaker was in a relaxed and normal emotional state. This test case serves as the basis for the remaining two test cases. All the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 11, November-2019

parameter statistics indicate mean values derived from the speech samples. As shown in Table I, for the purpose of demonstration, statistics for two speech samples have been analyzed.

TABLE I
AVERAGE VALUES OF VOCAL STATISTICS OBTAINED FROM RECORDED SPEECH SAMPLES FOR A NORMAL EMOTIONAL STATE

	Pitch (Hz)	SPL (dB)	Timbre ascend time (s)	Timbre descend time (s)	Time gaps between words (s)
Speech Sample 1	1248 Hz	Gain -50 dB	0.12 s	0.11 s	0.12 s
Speech Sample 2	1355 Hz	Gain -48 dB	0.06 s	0.05 s	0.12 s

- **Case 2**: *Angry* emotional state

This test case involves statistics for pitch, SPL, timbre, and word-timing gaps derived from speech samples that were orated while the speaker was in an agitated emotional state, typically characterized by increased vocal loudness and pitch. All the parameter statistics indicate mean values derived from the speech samples, as shown in Table II. The same speech samples that were earlier used in Case 1 have been used in Case 2, but with a different intonation typical of an agitated or angry emotional state.

TABLE II
AVERAGE VALUES OF VOCAL STATISTICS OBTAINED FROM RECORDED SPEECH SAMPLES FOR AN ANGRY EMOTIONAL STATE

	Pitch (Hz)	SPL (dB)	Timbre ascend time (s)	Timbre descend time (s)	Time gaps between words (s)
Speech Sample 1	1541 Hz	Gain – 30 dB	0.13 s	0.10 s	0.09 s
Speech Sample 2	1652 Hz	Gain – 29 dB	0.06 s	0.04 s	0.10 s

- **Case 3**: *Panicked* emotional state

This test case involves statistics for pitch, SPL, timbre, and word-timing gaps derived from speech samples that were orated while the speaker was in a panicked or overwhelmed emotional state. Speech samples that were earlier used in Case 1 have been used in Case 3, but with a different intonation typical of a panicked emotional state, as shown in Table III.

TABLE III
AVERAGE VALUES OF VOCAL STATISTICS OBTAINED FROM RECORDED SPEECH SAMPLES FOR A PANICKED EMOTIONAL STATE

	Pitch (Hz)	SPL (dB)	Timbre ascend time (s)	Timbre descend time (s)	Time gaps between words (s)
Speech Sample 1	1443 Hz	Gain – 46 dB	0.13 s	0.09 s	0.13 s
Speech Sample 2	1560 Hz	Gain – 44 dB	0.07 s	0.04 s	0.14 s

3. RESULTS ANALYSIS

The speech samples described in the previous section were recorded with the help of headphones that have an offset gain of approximately -60 dB and the speech data was sampled at the rate of ten sample points per second. By comparing Tables I and II, it can be seen that when in an agitated state, a significant increase occurs in the mean SPL and pitch values, accompanied by a decrease in the time spacing between consecutive words. In simple terms, this would indicate faster talking in a shrill and louder voice.

By comparing Tables I and III, it can be seen that when in a nervous or panicked state, there is a significant increase in the mean values of pitch, time spacing between consecutive words, and increased timbre ascending time. In simple terms, this would indicate a shrill voice with longer, sharp pauses.

By comparing the data presented in Tables 1-3, it can be decisively concluded that with varying emotions, the tonal parameters accordingly change as well. Establishing value-ranges for the aforementioned various vocal parameters can help in quantitatively assessing the extent of deviation from the standard basis, which in this study is the *normal* emotional state. The findings of the previous section are subjective and will vary depending on how a person reacts to a particular emotional situation. However, from the point of practicality, this feature of subjectivity can be exploited, especially while developing customizable applications, such as smart- phone apps that are user oriented.

Relevant objectives: -

1) This system is useful to detect either side of the speakers in telephone to identify the gender and based on it identifying age as well as its emotion.
2) There are many applications based on speaker's voice like, finding the gender, age group and providing the service according to the gender and age group.
3) Also include speaker recognition and displaying the speaker profile according to our database.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 11, November-2019

Motivation: -

- Very less research on human voice-based age and gender detection system.
- We can use this system according to market age group requirement and gender requirements.
- Recognize the emotions in the voice (sad, angry, happy etc.).
- Human voice can also be used as secured biometric characteristics. It helps to store criminal record using voice and can also be used as identify criminal's voice based on recorded audio file as an evidence.
- Also finding the gender and age automatically before the inquiry.

4. HYPOTHESIS:

Age detection system take human voice as input and find the gender and age of speaker and compare that voice with database profile audio file. If found the matched profile display the profile in detail, if not found display the gender and estimated age of speaker.

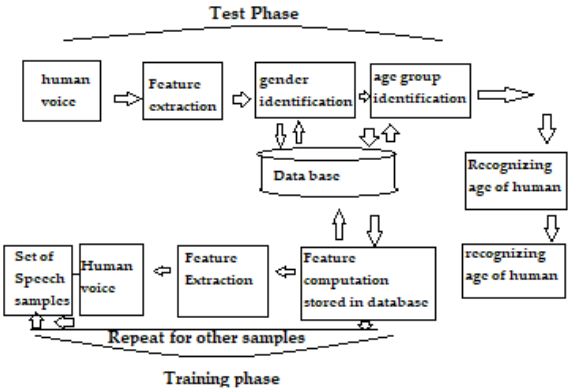

Fig1: Proposed system block diagram.

5. LITERATURE REVIEW:

In [1] Age and gender detection are two tasks, uses two separate technology to find the age and gender of human and use multiple technique to find the age and gender the calibrate and make fusion of that data and find the high greater result as compare to other technique.

In Speaker gender detection and age estimation, based on hybrid architecture of Weighted supervised Non-Negative Matrix Factorization (WSNMF) and General Regression Neural Network (GRNN). Here used the two phases to compare the data,

1) Training phase
2) Testing phase

Gaussian mixture weight super vectors of the primary training set are used to train a WSNMF which is applied for recognizing the age-gender category of any unseen speakers.

In [3] divide the human according to Age group and gender in seven classes and make database of speaker identify the speaker voice. And search the speaker profile with help of database, GMM/SVM-super vector system for speaker age and gender recognition, a technique that is adopted from state-of-the-art speaker recognition research.

6. PROBLEM STATEMENT:

It is very difficult to find the human age automatically. So, the proposed system will take a speaker voice as input and process it to find the gender, age group, exact age of speaker and his emotion state. The proposed system classifies the speaker age in group like children, young, adult, senior, then recognizing or compare the speaker profile with stored profile at proposed system. If profile found display profile of speaker with emotions otherwise display gender, age group and emotion.

7. SOLVING APPROACH:

There are multiple approaches to solve the problem. Many techniques are used like processing of voice data, feature extraction, Pattern recognition, voice recognition-based profile matching. According the above proposed system classify the sample data into two dataset training phase and testing phase, also consider the age and gender challenge separately to find the speaker characteristics efficiently.

Efficiency issues:

1) Speaker's Voice
2) Noisy environment
3) Variability in Voice
4) Depend on which Technique used
5) Emotions

Outcomes:

- Perform the gender classification (male or female).
- Classify the people according the age group.
- Display the speaker profile according to voice database with emotions.
- Display the emotions of speaker (angry, happy, sad etc.).

7. REFERENCES:

[1] Felix Burkhardt, Martin Eckert, Wiebke Johannsen, Joachim Stegmann ,"A Database of Age and Gender Annotated Telephone Speech ", Deutsche Telekom AG Laboratories, Ernst-Reuter-Platz 7, 10587 Berlin, Germany,2010.

[2] Michael Feld1, Felix Burkhardt2, Christian Muller "Automatic Speaker Age and Gender Recognition in the Car for Tailoring Dialog and Mobile Services" German Research Center for Artificial Intelligence.

[3] Hugo Meinedo1, Isabel Trancoso, "Age and Gender Classification using Fusion of Acoustic and Prosodic Features", Spoken Language Systems Lab, INESC-ID Lisboa, Portugal, Instituto Superior T´ecnico, Lisboa, Portugal,2008.

[4] Mohamad Hasan Bahari, Hugo Van hamme,"Speaker Age Estimation and Gender Detection Based on Supervised Non-Negative Matrix Factorization " , Centre for Processing Speech and Images Katholieke Universiteit Leuven,Leuven, Belgium.

[5] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. N¨oth, 'Age and Gender Recognition for Telephone Applications Based on GMM Supervectors and Support Vector Machines', in Proceedings of ICASSP 2008, Las Vegas, NV, (2008).

[6] L. Cerrato, M. Falcone, and A. Paoloni, 'Subjective age estimation of telephonic voice', Speech Communication, 31(2-3), 107–112, (2000).

[7] R. J. Davidse, M. P. Hagenzieker, P. C. van Wolffelaar, and W. H. Brouwer, 'Effects of in-car support on mental workload and driving performance of older drivers', Human Factors: The Journal of the Human Factors and Ergonomics Society, 51(4), 463–476, (2009).

[8] A. Gruenstein, J. Orszulak, S. Liu, S. Roberts, J. Zabel, B. Reimer, B. Mehler, S. Seneff, J. Glass, and J. Coughlin, 'City browser: developing a conversational automotive hmi', in Proceedings of the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 11, November-2019

27th international conference extended abstracts on Human factors in omputing systems, pp. 4291–4296, Boston, MA, USA, (2009). ACM New York, NY, USA.

[9] N. Minematsu, K. Yamauchi, and K. Hirose, 'Automatic Estimation of Perceptual Age Using Speaker Modeling Techniques', in Proceedings of Euro speech 2003), pp. 3005 – 3008, Geneva, Switzerland, (2003).

[10] P.H. Ptacek and E.K. Sander, 'Age recognition from voice', Journal of Speech and Hearing Research, 9, 273–277, (1966).

[11] S. Sch¨otz, Perception, Analysis and Synthesis of Speaker Age, Ph.D. dissertation, University of Lund, Sweden, 2006.

[12] S. Sch¨otz, 'Acoustic Analysis of Adult Speaker Age', in Speaker Classification, ed., Christian M¨uller, volume 4343 of Lecture Notes in Computer Science / Artificial Intelligence, Springer, Heidelberg - Berlin - New York, (2007). this issue.

[13] J. M. Wood, 'Aging, driving and vision', Clinical and Experimental Optometry, 85(4), 214–220, (2002).

[14] E. Shriberg, L. Ferrer, S. Kajarekar, N. Scheffer, A. Stolcke, and M. Akbacak, "Detecting nonnative speech using speaker recognition approaches", in Proceedings IEEE Odyssey-08 Speaker and Language Recognition Workshop, Stellenbosch, South Africa, Jan. 2008.

[15] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation", in Proceedings of EMNLP, 2010.

[16] "Census regions and divisions of the united states", http://www.census.gov/geo/www/us regdiv.pdf.

[17] D.M. Bikel, S.Miller, R. Schwartz, and R.Weischedel, "Nymble: a high-performance learning name-finder", in In Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 194–201, 1997.

[18] W. Wang, G. Tur, J. Zheng, and N. F. Ayan, "Automatic disfluency removal for improving spoken language translation", in Proc. ICASSP, 2010.

[19] W. Wang, "Weakly supervised training for parsing Mandarin broadcast transcripts", in Proceedings of Interspeech, pp. 2446–2449, Brisbane, Australia, September 2008.

[20] C. Whissell, "Whissell's dictionary of affect in language: Technical manual and user's guide", Laurentian University, http://www.hdcus.com/manuals/wdalman.pdf.

[21] "Linguistic inquiry and word count", http://www.liwc.net/liwcdescription.php.

[22] A. Stolcke, S. Kajarekar, and L. Ferrer, "Nonparametric feature ormalization for SVM-based speaker verification", in Proc. ICASSP, pp. 1577–1580, Las Vegas, Apr. 2008.